Capturing, Analyzing, and Transmitting Intangible Cultural Heritage with the i-Treasures Project

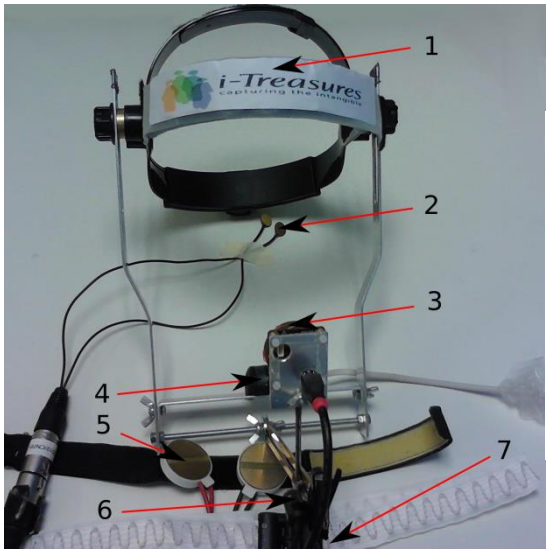
A. Jaumard-Hakoun[1,2], S. K. Al Kork[1,2], M. Adda-Decker[3], A. Amelot[3], L. Buchman[3], G. Dreyfus[2] , T. Fux[3], P. Roussel[2], C. Pillot-Loiseau[3], M. Stone[4], B. Denby[1,2]

[1]Université Pierre et Marie Curie, Paris, France
[2]Signal Processing and Machine Learning Lab, ESPCI-ParisTech, Paris, France
[3]Phonetics and Phonology Lab, CNRS UMR 7018, University Paris 3 Sorbonne Nouvelle
[4]Vocal Tract Visualization Lab, University of Maryland Dental School, Baltimore, USA

I. Project overview

The i-Treasures project, which officially began on 1 February 2013, is a 12-partner FP7 project that proposes to use multi-sensor technology to capture, preserve, and transmit four types of intangible cultural heritage, referred to as 'use cases': rare traditional songs, rare dance interactions, traditional craftsmanship and contemporary music composition. Methodologies used will include body and gesture recognition, vocal tract modeling, speech processing and electroencephalography (EEG). The "Rare traditional songs" use case, which will be the focus of our work, targets Corsican "cantu in paghjella", Sardinian "canto a tenore", Mt. Athos Greek byzantine hymns and the recent "Human beat box" styles. The final objective of the "Rare traditional songs" use case is to capture vocal tract movements with sufficient accuracy to drive a real-time 2D or 3D avatar of the vocal tract, which will in turn play a crucial role in the transmitting of captured invisible cultural heritage to future generations.

II. Acquisition system



1. Multi-sensor helmet
2. Nose-mounted accelerometer
3. Camera
4. Ultrasound probe
5. Electroglottograph
6. Microphone
7. Breathing belt

Figure 1. Acquisition sensors.

To study vocal tract dynamics during singing performances, artists will be instrumented with a non-intrusive acquisition helmet (see figure 1) containing an ultrasound probe, a camera and a microphone. Additional sensors will complete the acquisition system: a piezoelectric accelerometer mounted on the nose, an Electroglottograph (EGG) necklace and a breathing belt. The ultrasound probe will be used to study tongue movements, articulation and lingual contour [1]. The camera will help us detecting lips and jaw movements, lips aperture and protrusion. The microphone provides an acoustic reference of sound production and enables spectral analysis. Nasality will be measured thanks to the accelerometer while the EGG will provide information concerning glottis dynamics. The kind of breathing as well as breathing rhythm (amplitude, frequency) will be explored through the breathing belt. Information concerning vocal quality will be extracted from speech, EGG and accelerometer signals [2, 3]. Both ultrasound probe and camera provide grayscale pictures of size respectively 320 x 240 and 640 x 480 pixels acquired at 60 fps. Signals from the other sensors are mono-dimensional and sampled at 44100 Hz. The proposed acquisition system is developed using a Real Time Modular Application System (RTMaps) [4] and provides real-time data displaying and recording, which can be either local or on a network.

III. Vocal tract modeling

Capturing and modeling intangible cultural heritage is a challenging task that will require appropriate sets of salient descriptors derived from sensor data. Currently, data-driven "deep learning" or DL architectures are being investigated as a means of extracting articulatorily pertinent vocal tract information from ultrasound of the tongue and lips images. The DL approach, which has been applied to pattern or phoneme recognition, is one of the most efficient unsupervised learning methods. A typical DL architecture consists of a neural network with a large number of hidden layers, with each additional layer corresponding to a higher level of abstraction, a hierarchy which imitates the layered structure of the brain's visual and auditory cortexes. Pre-training such a network, in an unsupervised way instead of initializing it randomly, provides both optimization and regularization of the network and reduces training error [5].

IV. Preliminary data processing

Such DL architectures applied on ultrasound images may be relevant to extracting salient descriptors from articulatory data and modeling vocal tract movements during singing performance. We aim at extracting the contour of the tongue from ultrasound images in a speaker-independent manner. If we train a network on a large database of several singers' recordings, the intuition is that our network will be able to extract descriptors regardless the configuration of the mouth, provided the network is sufficiently trained. Another reason for using DL is that once the network is optimized for a specified task, it is easy to use it to perform this task without any human supervision. Processing one image only requires one single pass through the network, which allows real-time processing. Assuming we have a contour for a subset of our ultrasound data (manually or automatically extracted), our network is trained on a representative learning database composed for each example of a reduced image of ultrasound data and a contour image. Our input contour images are obtained thanks to an automatic tongue contour extraction algorithm: for each columns of each image, several pixels are selected as contour candidates. Decision is made according to the shape of the previous image and the distance between the candidate and the last point detected as contour. In our work, as shown in figure 2, we first reduce the dimensions of input images (both ultrasound and contour images are reduced to 30x33 pixels) so that the number of neurons in the input layer of our neural network is not too large.

V. Extracting tongue contour with deep networks

We used an autoencoder structure (encoder followed by a decoder, respectively the lower and the upper part of the network, see figure 2) so that the network is trained to reconstruct the input from the descriptors of its hidden layers, which contains an abstract representation of the inputs. The construction of this autoencoder is made through greedy layer-wise Restricted Boltzmann Machines (RBM) training [6]. A RBM is a neural network made of a visible unit and a hidden unit defined by its probability distribution (Bernoulli). The propagation rule is given by joint probabilities and weights are updated according to a learning rule. A deep network is made of stacked and sequentially-trained RBM, so that the hidden layer of each RBM

is the visible unit of the following RBM. Hidden layers contain a compressed representation of input data that seems sufficient to reconstruct a contour from sensor-only input data using "translational" RBM defined in [7]. The idea is to train a network on both ultrasound and contour images as input to learn a representation of these inputs and their relationship. Given these shared features, the decoder stage is able to reconstruct both images. Then if a new encoder can be trained on ultrasound data only to produce hidden features, identical to those produced by the network with concatenated ultrasound and contour data, it is possible to reconstruct labels from hidden features using the decoder of the original network (see figure 3).

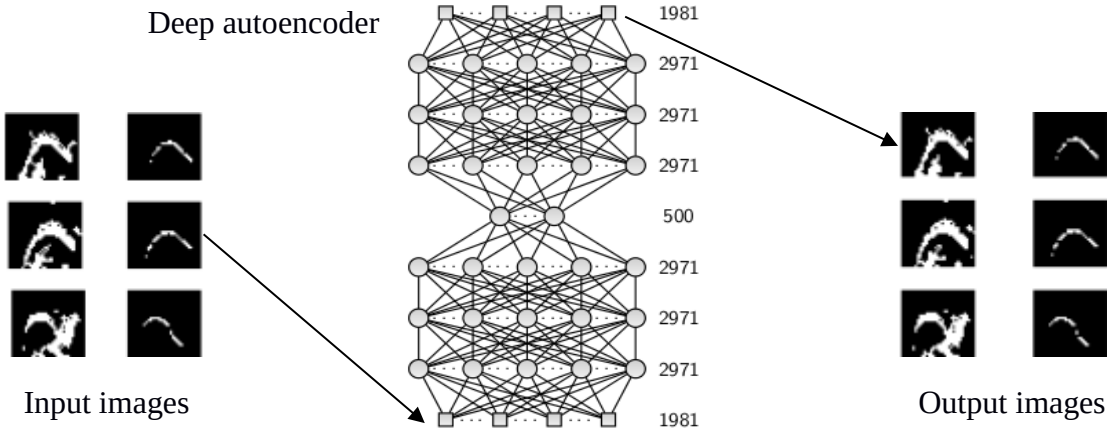


Figure 2. Deep Autoencoder training.

The first layer of our network is composed of one neuron per pixel of the ultrasound image and one neuron per pixel of the contour image plus one for bias, that is 1981 neurons (see figure 2). Our network is composed of 7 hidden layers of various lengths (figure 2 and 3). Once our deep autoencoder is trained to reconstruct input images (figure 2), we use our translational network on new ultrasound images to reconstruct both ultrasound and contour images (figure 3). Despite the low quality of input images, we are able to convert output images into a set of coordinates that represents the contour of the tongue. In figure 4, red points represent the output of the translational network after conversion into pixel coordinates, which combines contour image thresholding, isolated points removal and contour interpolation. However, the quality of reconstructed contour can be very poor if the input image is too noisy and quite sensitive to artifacts. Improving input image quality can be considered as one option, although it would dramatically increase computing time.

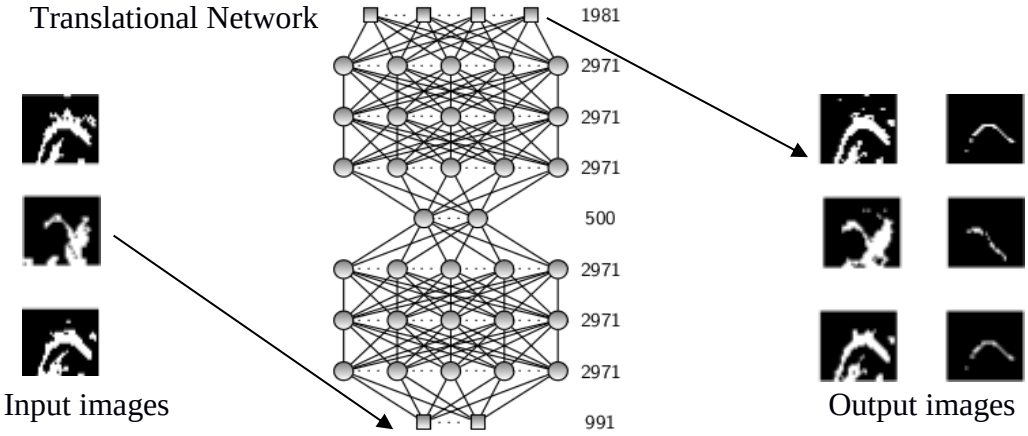


Figure 3. Translational network use.

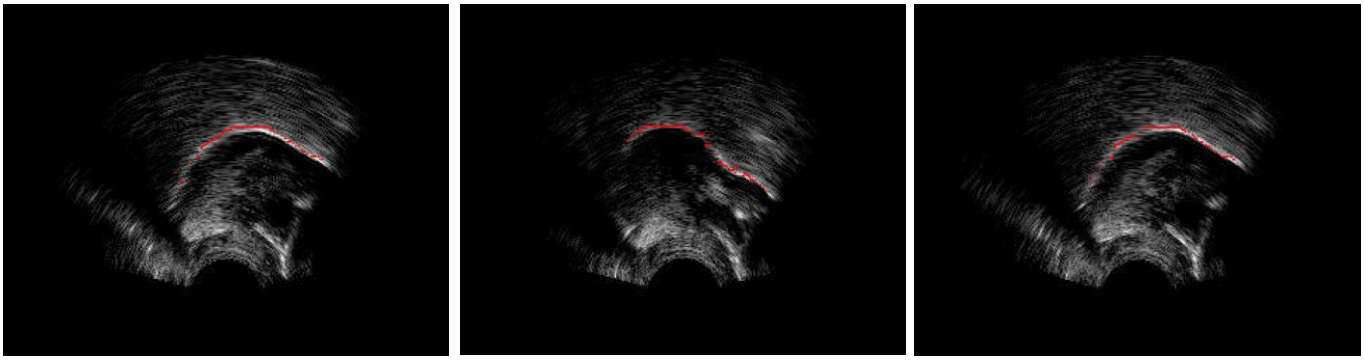


Figure 4. Examples of tongue contour extraction results.

VI. Conclusions and future work

Using a DL approach is a useful way to extract salient descriptors from ultrasound images. The advantage of our method is the use of automatically extracted contours as inputs to train our neural network so that the network is able to learn tongue contour extraction. We will extend this method to lip images processing. In addition, since our current sensor processing algorithms are offline tests, our next steps will concern the integration of a sensor output processing block to our current real-time module.

VII. Acknowledgements

This work is funded by the European Commission via the i-Treasures project (Intangible Treasures - Capturing the Intangible Cultural Heritage and Learning the Rare Know-How of Living Human Treasures FP7-ICT-2011-9-600676-i-Treasures). It is an Integrated Project (IP) of the European Union's 7th Framework Program 'ICT for Access to Cultural Resources'.

VIII. References

[1] M. Stone, *A Guide to Analyzing Tongue Motion from Ultrasound Images,* Clinical Linguistics and Phonetics, 19(6-7), 455-502, 2005.

[2] N. Henrich, C. D'Alessandro, B. Doval, M. Castellengo, *Glottal open quotient in singing: Measurements and correlation with laryngeal mechanisms, vocal intensity, and fundamental frequency*, The Journal of the Acoustical Society of America 117(3), 1417-1430, 2005.

[3] C. D'Alessandro, *Method in empirical prosody research*, chapter "Voice Source Parameters and Prosodic Analysis", 63–87, 2006.

[4] INTEMPORA S.A., 2011. [Online]. Available: http://www.intempora.com/.

[5] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent and S. Bengio, *Why Does Unsupervised Pre-training Help Deep Learning ?,* J. Mach. Learn. Res., 2010.

[6] G. Hinton, S. Osindero, and Y. The, *A fast learning algorithm for deep belief nets*, Neural Computation, 18(7),1527–1554, 2006.

[7] I. Fasel and J. Berry, *Deep Belief Networks for Real-Time Extraction of Tongue Contours from Ultrasound During Speech*, ICPR, 1493-1496. IEEE, 2010.